Anatomically Consistent CNN-Based Segmentation of Organs-at-Risk in Cranial Radiotherapy

Pawel Mlynarski[a,1], Hervé Delingette[a], Hamza Alghamdi[b], Pierre-Yves Bondiau[b], Nicholas Ayache[a]

[a] *Université Côte d'Azur, Inria, Epione research team, France.*
[b] *Université Côte d'Azur, Centre Antoine Lacassagne, France.*

Abstract

Planning of radiotherapy involves accurate segmentation of a large number of organs at risk, i.e. organs for which irradiation doses should be minimized to avoid important side effects of the therapy. We propose a deep learning method for segmentation of organs at risk inside the head, from Magnetic Resonance (MR) images. Our system performs segmentation of eight structures: eye, lens, optic nerve, optic chiasm, pituitary gland, hippocampus, brainstem and brain. We propose an efficient algorithm to train neural networks for an end-to-end segmentation of multiple and non-exclusive classes, addressing problems related to computational costs and missing ground truth segmentations for a subset of classes. We enforce anatomical consistency of the result in a postprocessing step, in particular we introduce a graph-based algorithm for segmentation of the optic nerves, enforcing the connectivity between the eyes and the optic chiasm. We report cross-validated quantitative results on a database of 44 contrast-enhanced T1-weighted MRIs with provided segmentations of the considered organs at risk, which were originally used for radiotherapy planning. In addition, the segmentations produced by our model on an independent test set of 50 MRIs are evaluated by an experienced radiotherapist in order to qualitatively assess their accuracy. The mean distances between produced segmentations and the ground truth ranged from 0.1 mm to 0.7 mm across different organs. A vast majority (96 %) of the produced segmentations were found acceptable for radiotherapy planning.

1. Introduction and related work

Malignant tumors of the central nervous system cause more than 200 000 deaths per year worldwide [48]. Many brain cancers are treated with radiotherapy, often combined with other types of treatment, in particular surgery and chemotherapy. Radiotherapy planning requires segmentation of target volumes (visible tumor mass and regions likely to contain tumor cells) and anatomical structures that are susceptible to be damaged by ionizing radiation exposure during treatment. The segmented volumes are used for computation of optimal irradiation doses, with the objective of maximizng irradiation of cancer cells while minimizing damage of neighboring healthy structures, called *organs at risk* (OAR). Magnetic Resonance (MR) images [4] are commonly used for imaging of tumors and organs in the head. In this work, we address the challenging problem of multiclass segmentation of organs in MRI of the brain.

Delineation of organs at risk is today manually performed by experienced clinicians. Due to a large number of structures to be accurately segmented, the segmentation process takes usually several hours per patient. Manual

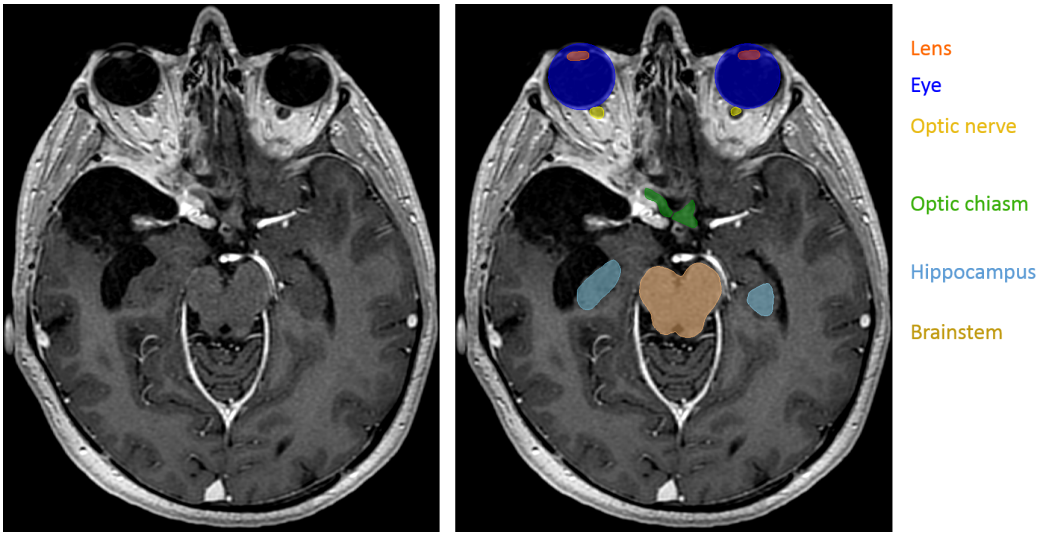

Figure 1: Segmentation of organs at risk in radiotherapy planning. Left: T1-weighted MRI acquired after injection of a gadolinium-based contrast agent. Right: manual annotations of several organs at risk. In contrast to standard segmentations problems, one voxel may belong to zero or several classes (for instance, the eye and the lens).

segmentation represents therefore a very high cost and eventually delays the beginning of the therapy. Moreover, a high intra-observer and inter-observer variability is observed [9]. Automatic methods for segmentation of organs at risk are therefore of particular interest. We can distinguish two main types of approaches proposed in the literature.

The first type of methods corresponds to atlas-based approaches [10, 7, 1]. The input image is typically registered to one [12, 13] or several [41, 42] annotated images, from which the segmentation is extrapolated. When multiple atlases are used, the candidate segmentations may be combined, for instance, by voting strategies [42] or by the STAPLE algorithm [51]. An important advantage of atlas-based methods is to produce anatomically consistent results. However, their main drawback is their limited generalization capacity. The important variability between cases results not only from the natural anatomical differences between patients but also from pathological factors. In particular, healthy organs are deformed by growing tumors, which may appear at different locations and which are typically not present in atlases. Some organs may even be missing because of surgeries undergone previously by the patient.
The second group of approaches is based on a discriminative classification of voxels with machine learning models such as Random Forests [16, 15, 22] or Convolutional Neural Networks (CNN) [29]. These discriminative methods are less constrained than atlas-based approaches and may therefore better adapt to the diversity of cases. However, in general, voxelwise classifiers may produce results which are inconsistent in terms of shapes and locations of organs.

Organs at risk in the head have complex shapes and are surrounded by other structures sharing similar voxel intensities in MRI. Moreover, there are large differences related to acquisition of MRI, especially when images come from different medical centers. In order to segment organs from MRI, a complex and abstract information has therefore to be extracted. Convolutional Neural Networks are suitable for this task, as they have the ability to automatically learn complex and relevant image features. In this work, we propose a system based on CNNs for multiclass segmentation of organs at risk in brain MRI.

In this work we assume non-exclusive classes, i.e. that one voxel may belong to zero or several classes (Fig. 1). This is in contrast with the majority

of segmentation models, which assign one unique label to each voxel following the format of public segmentation challenges such as the BRATS [32]. However, some works addressing OAR segmentation consider non-exclusive classes [37, 50, 27, 23] similarly to our work. An important difficulty to train machine learning models for multiclass OAR segmentation is the varying availability of ground truth segmentations of different classes among patients, depending on clinical needs. While some organs, such as the optic nerve, are systematically segmented during radiotherapy planning, annotation of other structures may be available only for a subset of patients. One solution to this problem is to independently train one model per class, as it was proposed in some recent deep learning works [27, 23, 31]. A limitation of this approach is, however, the need to perform time-consuming trainings for every class, while the number of classes of interest may be large. In this work, we propose a loss function and an algorithm to train neural networks for an end-to-end multiclass segmentation, taking into account the problem of missing annotations. To the best of our knowledge, the only deep learning method for end-to-end multiclass OAR segmentation which addresses this issue is the one proposed in [54] for the segmentation of head and neck organs at risk in CT scans.

The network architecture used in our work is a modified version of 2D U-net [43]. As computation of gradients of the loss function by Backpropagation represents high computational costs and GPU memory load, large segmentation CNNs often cannot be trained on entire MRIs or CT scans (representing several millions of voxels). Input images are often downsampled [44, 50] to allow large 3D CNNs to capture information from distant regions, i.e. to have a long-range 3D receptive field. In this work, we have chosen a 2D architecture to limit the memory load while capturing a long-range spatial context without the need to downsample input images. In particular, downsampling of inputs may affect segmentation of small structures such as the optic nerve or the lens..

Even if most of the proposed deep learning methods for OAR segmentation do not apply anatomical constraints on the output of neural networks, some approaches include shape priors in models. For instance, [47] propose to learn latent representations of shapes of organs by a stacked autoencoder and to use these learned representations in the loss function of a segmentation network, in order to compare the shape of the output with the shape of the ground truth. The works [8, 38] propose to adapt triangulated meshes representing organ boundaries to medical images and to use neural networks

for regression of distances between centers of triangles and organ boundaries. This type of approach may therefore be seen as atlas-based with the use of deep learning for boundary detection.
Spatial relations between anatomical structures in the head could be explicitly modeled using, for instance, graph-based representations as it was proposed in model-based methods [6, 21]. However, inclusion of constraints related to connectivity and relative positions of organs in loss functions of CNNs is not trivial due to considerable computational costs. In order to apply such constraints, a neural network would have to segment large regions of the input images during the training phase. Segmentation of large 3D volumes containing different anatomical structures requires a considerable amount of the GPU memory, as outputs of all layers of the network are stored in the GPU during computation of gradients of the loss function. Moreover, penalization of anatomical inconsistencies during the training does not guarantee anatomically consistent results at the test phase. To the best of our knowledge, none of the proposed deep learning methods explicitely enforces consistency of OAR segmentation in terms of relative positions of organs. However, some methods define regions of interest of organs, for instance by registering the image to a set of atlases [27].

In our work, we enforce some anatomical constraints in a postprocessing stage, starting from the segmentation produced by majority voting of 2D CNNs processing the image by axial, coronal and sagittal slices. In particular, we propose an anatomically consistent segmentation of the optic nerves, with an approach based on the search of the shortest path in a graph, using outputs of neural networks to define weights of edges in the graph. Application of anatomical constraints in postprocessing modules rather than in the deep learning model is mainly motivated by computational costs of CNNs but also by their 'black box' aspect.

We consider eight classes of interest, corresponding to anatomical structures systematically segmented during radiotherapy planning for brain cancers: eye, lens, optic nerve, optic chiasm, pituitary gland, hippocampus, brainstem and brain (including cerebrum, cerebellum and brainstem. The anatomical structures composed of left and right components (eye, lens, optic nerve, hippocampus) are seen as one entity by the neural network but are separated in the postprocessing step.

Most of the proposed deep learning methods for segmentation of organs at risk were applied on CT scans in the context of head and neck cancers [3],

i.e. cancers of the upper parts of respiratory and digestive systems (mouth, larynx, throat). To the best of our knowledge, the only deep learning method for segmentation of organs at risk in MRIs of the brain is the one proposed in [38] (MRI T1 and T2).

Our method is tested on a set of contrast-enhanced T1-weighted MRIs acquired in the Centre Antoine Lacassagne in Nice (France). First, our method is quantitatively evaluated on a set of 44 MRIs with provided segmentation of different anatomical structures. Segmentation performances are measured by three different metrics: Dice score, Hausdorff distance and the mean distance between the output and the ground truth. Then, the segmentations produced by our method on a different set of 50 MRIs are qualitatively evaluated by an experienced radiotherapist. Our system was able to produce segmentations with an accuracy level which was found acceptable for radiotherapy planning in a large majority of cases (96%). The mean distances between the output segmentation and the ground truth for different organs were between 0.1 mm and 0.7 mm.

2. Methods

2.1. Deep learning model

2.1.1. Network architecture

The architecture used in our work is a modified version of 2D U-Net [43], which is composed of an encoding part and a decoding part. The encoding part is a sequence of convolutional and max-pooling layers. The number of feature maps is doubled after each pooling, taking advantage of their reduced dimensions. The decoding part is composed of convolutional and upsampling layers. Feature maps of the encoding part as concatenated in the decoding part in order to combine low-level and high-level features and to ease the flow of gradients during the optimization process. The final convolutional layer (the segmentation layer) of the standard U-Net has two feature maps, representing pixelwise classification scores of the class 0 ('background') and the class 1. During training, these two final feature maps are normalized by the softmax function.

We adapt this architecture to our problem of multiclass segmentation with non-exclusive classes, where each pixel may belong to zero or several classes. In the following, C denotes the number of classes (in our experiments, $C = 8$) and the classes are numbered from 1 to C. In our model, each class c has

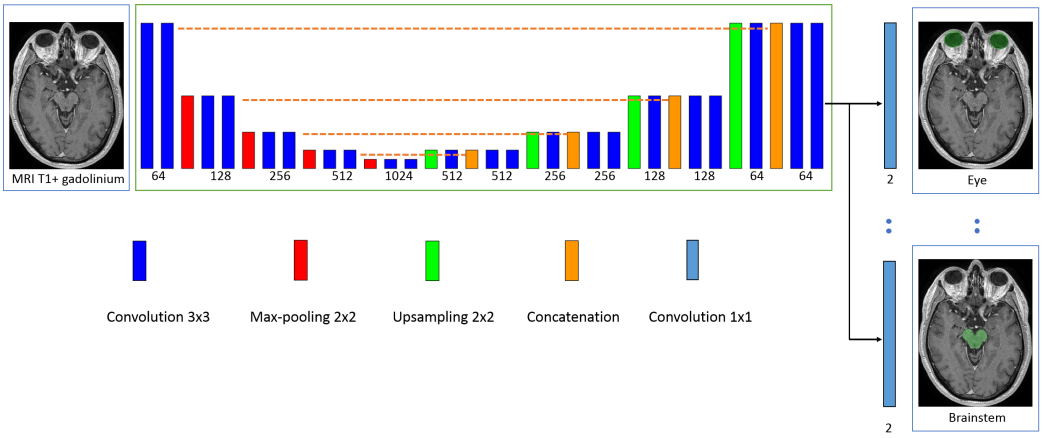

Figure 2: Architecture of our model. The rectangles represent layers and their height represents the sampling factor (increasing with max-poolings, decreasing with upsamplings). The numbers of features maps are specified below layers. The proposed model is a modified version of U-Net, having one segmentation layer per class in order to perform an end-to-end multiclass segmentation with non-exclusive classes.

its dedicated binary segmentation layer (Fig. 2), composed of two feature maps corresponding to pixelwise scores of the class and of the background. Each segmentation layer takes as input the second to last convolutional layer of U-Net. We use batch normalization [24] in all convolutional layers of the network, except segmentation layers.

2.1.2. Training of the model

Our loss function and training scheme were designed to deal with class imbalance and the problem of missing annotations (for a given image, the ground truth is available only for a subset of classes).
In a given training image i, each pixel (x,y) has 3 possible labels for the class c: 0 (negative), 1 (positive) or -1 (unknown). If the ground truth segmentation of the class c is unavailable for the image i, all pixels are labelled as unknown for the class c by default. However, missing annotations may be partially reconstructed from segmentations of other classes. For example, if the segmentation of the 'lens' class is not available but the 'eye' class is segmented, all pixels outside the eye may be labelled as negative for the lens.

Given a training batch of M images and the estimated parameters θ of the network, the segmentation layer of the class c is penalized by the following

loss function, which can be seen as pixelwise cross-entropy with adaptative weights. Let's note N_0^c, N_1^c and N_{-1}^c the numbers of pixels labelled respectively 0, 1 and -1 for the class c in the training batch. The weight $w_{(x,y)}^i$ of the pixel (x,y) of the image i has three possible values, according to the label of the pixel. If the label is unknown, then $w_{(x,y)}^i = 0$. If its label is 1, then $w_{(x,y)}^i = t_c/N_1^c$ where $0 < t_c < 1$ is a fixed hyperparameter, which we call the *target weight*. If the pixel is labelled 0, then $w_{(x,y)}^i = (1 - t_c)/N_0^c$. The introduced hyperparameter t_c controls therefore the relative weight of positive and negative pixels of the class c (positive pixels have the total weight of t_c and negative pixels have the total weight of $1 - t_c$). This type of weighting strategy has been used in our previous work [34] to counter the problem of class imbalance. The loss function of the segmentation layer of the class c is defined by $Loss_c(\theta) = -\sum_{i=1}^{M} \sum_{(x,y)} w_{(x,y)}^i \log(p_{i,(x,y)}^l(\theta))$ where $p_{i,(x,y)}^l$ is the softmax score given by the network for the ground truth label l of the pixel. The loss function of the model is a convex combination of losses of all segmentation layers: $Loss(\theta) = (1/C) \sum_{c=1}^{C} Loss_c(\theta)$.

We propose a sampling strategy to construct training batches so that there are positive and negative pixels for each of the C classes in each training batch.

For each image of the training database, we precompute bounding boxes of all classes with provided segmentations. For bilateral classes such as the eyes, there are generally two bounding boxes per image corresponding to left and right components, unless one of the components is missing (e.g. an organ removed by surgery). The precomputed bounding boxes are used during the training in order to sample patches containing positive pixels of different classes.

At the beginning of the training, for each class c, we construct a list I_c of training images with provided ground truth segmentation of the class c. To sample a 2D patch which is likely to contain positive pixels of the class c, we randomly choose an image i from I_c and a random point (x, y, z) from the bounding box (or two bounding boxes if the class has left and right components) of the class c in the chosen image. Once the point is chosen, a 2D patch centered on this point is extracted from the image i and segmentations of all available classes are read. In the following, we refer to this procedure as extracting a patch centered on the class c.

We assume that the number of images in each training batch (M) is larger than the number of classes C, in order to be able to sample at least one image/patch centered on each of the classes. Each training batch is constructed as follows. The first C images of the batch are centered respectively on each of the C classes. At this stage, the batch is likely to contain positive and negative pixels of each class. The remaining $M - C$ images may be chosen randomly or be centered on larger classes. In our case, $C = 8$, $M = 10$ and the last images are centered on the largest class we segment, the brain, whose bounding box occupies almost an entire volume of the head.

As the model is trained for multiclass segmentation with non-exclusive classes, several binary segmentation maps have to be read for each input image in each iteration of the training. If the ground truth segmentations are not optimally stored in the memory, these reading operations may considerably slow down the training. The ground truth label of a given pixel can be represented by one bit (0 or 1). However, to store binary segmentation masks in commonly used formats such as HDF5 [20], each label would have to be represented by at least one byte. We propose therefore to store multiclass segmentations in a specifically encoded format, where every bit represents a label of a given class c. A binary segmentation mask of the class c is retrieved by the bitwise 'and' operation between the encoded multiclass segmentation and the code of the class, corresponding to a power of 2.
The size of extracted 2D patches should be chosen according to the capacities of the GPU. In our experiments, the training batches were composed of 10 patches of size 230x230. Given that in our network we use unpadded operations (convolutions, max-poolings, etc.), the dimensions of the outputs of segmentation layers are considerably smaller.
The model is trained with a variant of Stochastic Gradient Descent with momentum presented in our previous work [34]. The main characteristics of this algorithm is that gradients are computed over several batches in each iteration of the training, in order to use many training examples despite GPU memory limitations.

2.2. Postprocessing and enforcing anatomical consistency

Fully-convolutional neural networks such as our model produce segmentations by individually classifying every voxel based on intensities of voxels within the corresponding receptive field. Such classification is performed by extracting powerful and automatically learned image features. However, as

this classification is performed on a voxel by voxel basis, there is no guarantee of obtaining an anatomically consistent result, especially when the number of training images is limited. In particular, CNNs do not explicitely take into account aspects such as relative positions of different structures or adjacency of voxels belonging to the same structure. Including constraints related to these aspects in loss functions of neural networks or conceiving architectures which produce anatomically consistent results is difficult, in particular because of computational costs (need to simulatenously segment large 3D regions of input images). We propose therefore to improve consistency of segmentations in a postprocessing step. We also separate left and right components of classes such as the eye, as these components are considered separately for radiotherapy planning.
We combine, by majority voting, segmentations produced by three networks trained respectively on axial, coronal and sagittal slices. The goal of this combination is to take into account the three dimensions and to improve the robustness of the method. We subsequently apply a few rules described in the following, in order to correct some observed inconsistencies.

2.2.1. Segmentation of the brain

Brain (including the cerebrum, the cerebellum and the brainstem) is the largest class to be segmented. For various reasons, some voxels within this structure may be inconsistently classified as negative by networks, which appears as 'holes' in the segmentation or unrealistically sharp borders. We propose therefore a procedure that we call triplar hole-filling (Fig. 3). For each axial, coronal and sagittal plan of the 3D segmentation, we compute connected components of the background (negative voxels) and we remove components (changing their label from 0 to 1) which are not connected to the border of the plan. The reason of applying this procedure in 2D is that some holes may easily be connected to the outside of the class in 3D.

The bounding box of the segmentation of the brain is subsequently used to separate left and right components of bilateral classes. Note that the head of the patient may appear at different locations of the image, depending on acquisition conditions and performed preprocessings. For a given class expected to have left and right components (eye, lens, optic nerve, hippocampus), barycenter of each connected component is computed. In order

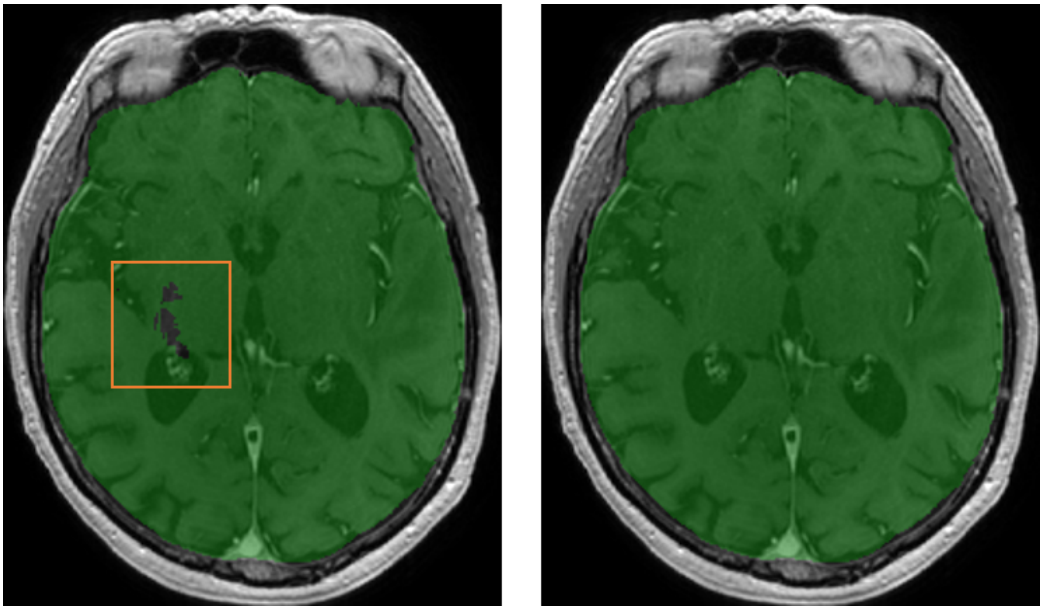

Figure 3: Example of 'holes' in the original output segmentation (left image) on a test example. Right image: segmentation obtained after our postprocessing (triplanar hole-filling).

to decide to which side corresponds a connected component, the coordinate x (right-left) of its barycenter is compared to min and max coordinates x of the bounding box of the brain.

2.2.2. Segmentation of the visual system

We propose an anatomically consistent segmentation of the visual system (eyes, lenses, optic nerves and chiasm), starting from the segmentations predicted by neural networks.

The eye is probably the less challenging organ for automatic segmentation as it has a simple spherical shape. However, some false positives are possible, especially in cases where an eye has been removed by surgery, resulting in false positives within the orbit. We propose therefore to remove connected components of eye segmentation whose volume is below an expected minimum value, which is set to 4 cm^3.

We constraint segmentation of the lenses to be inside the eyes, i.e. we assign the 0 label to all voxels outside the predicted masks of the eyes. Segmentation of the optic chiasm is obtained by taking the largest connected component

of the segmentation predicted by the networks. We distinguish left and right sides of the chiasm in order to compute landmarks for segmentation of the two optic nerves as described in the following.
Segmentation of the optic nerve in MR images is particularly challenging as the nerve is thin and may have an appearance similar to neighboring structures at some locations. However it has a rather regular shape which can be seen as a tube connecting an eye and the optic chiasm. The nerve is generally well visible at some locations, in particular close to the eye. A human expert is able to track the trajectory of the nerve to distinguish it from neighboring structures at more difficult locations. Based on this observation, we propose a graph-based algorihm for segmentation of the optic nerves in order to guarantee connectivity between the eyes and the optic chiasm and to decrease the number of false positives. The algorithm is based on the search of the shortest path between two nodes in a graph. Outputs of neural networks are used to define weights of the edges in the graph. The different steps of the algorithm (applied separately for left and right nerves) are described below.

First, we detect landmarks corresponding to the two endpoints of an optic nerve based on the initial segmentation of the visual system produced by neural networks (Fig. 4). The first landmark of the left optic nerve is the barycenter of P points initially predicted as the left optic nerve and which are closest to the left eye. The second landmark is computed similarly but searching P points of the left side of the optic chiasm which are the closest to the initial prediction of the left optic nerve. We take the barycenter of several points (in our experiments $P = 30$) in order to obtain a point which is more likely to be close to the centerline of the nerve. If the detected chiasm landmarks for the two optic nerves are anormally close, the procedure is applied only for one nerve, connecting the landmark with the closest eye.

Before applying the graph-based algorithm, we refine the initial segmentation of the optic nerves based on voxel intensities (specific to each image). In fact, the optic nerves are surrounded by fat, which appears hyperintense on MR T1-weighted images and can be rather easily distinguished from the optic nerve. We compute an approximate range of intensities of voxels of the fat by computing the 98% quantile of a small volume surrounding the eye-nerve landmark. Voxels whose intensites are above 80% of this value are classified negative for the optic nerve in order to eliminate common false

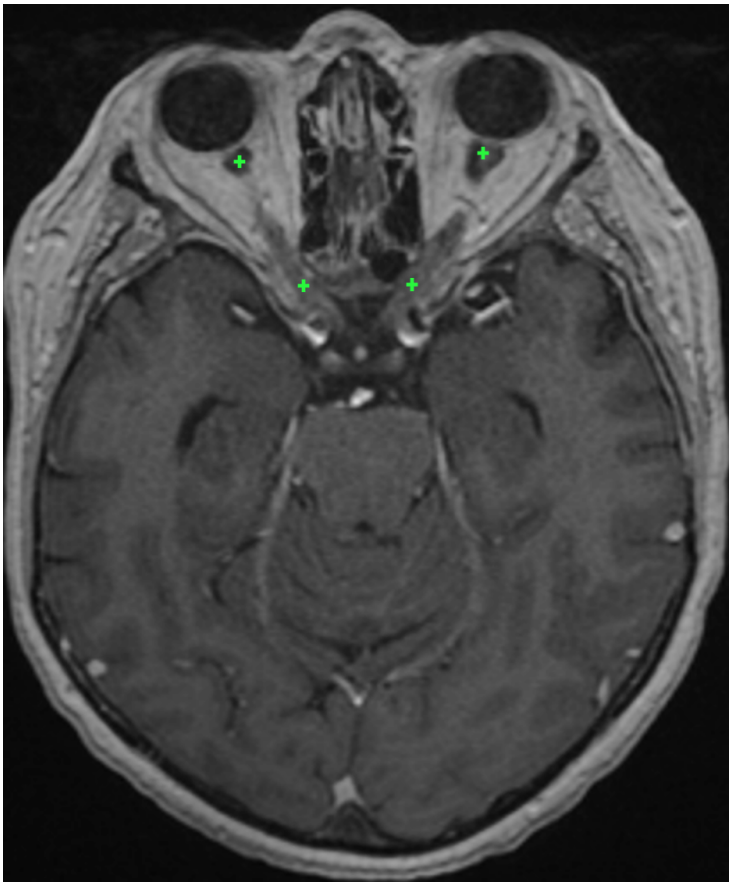

Figure 4: Approximate position of the optic nerve landmarks (displayed on the same axial slice) found by the system on a test example. For each of the optic nerves, the graph-based algorithm ensures the connectivity between the two landmarks.

positives.

Given the two computed landmarks and the refined inital segmentation, we estimate the centerline of the optic nerve (Fig. 5) by computing the shortest path in an oriented graph. The nodes of the graph correspond to voxels within a region of interest (cuboid containing the two landmarks) and which are reachable from the starting point. The connectivty of nodes is defined by adjacence of voxels with increasing y coordinate, i.e. the childs of the node (x, y, z) are nodes $(x + d_x, y + 1, z + d_z)$ with $d_x \in \{-1, 0, 1\}$ and $d_z \in \{-1, 0, 1\}$. We therefore assume strictly increasing y of the centerline

towards the second landmark (from anterior to posterior).
Each node (x, y, z) of the graph has its associated cost based on three criteria (listed by decreasing importance):

- Label $l_{(x,y,z)}$ initially assigned to the voxel (x, y, z). A strong penalty is applied to voxels predicted as negative, in order to force the centerline to pass by points initially predicted as positive. The associated cost is $c^{label}_{(x,y,z)} = 0$ if $l_{(x,y,z)} = 1$ and $c^{label}_{(x,y,z)} = C^l$ otherwise, where C^l is a fixed number controlling the importance of this cost (we set $C^l = 100$).

- If the predicted label $l_{(x,y,z)}$ is positive: distance d_{border} to the closest point classified as negative. The penalty is inversely proportional to this distance, to give priority to points which are far from predicted borders of the optic nerve (preference to central points). This cost is expressed by $c^{border}_{(x,y,z)} = 0$ if $l_{(x,y,z)} = 0$ and $c^{border}_{(x,y,z)} = R - d_{border}$ otherwise, where R is the radius of a search zone around the voxel (x, y, z). As the visible nerve is larger close to the eye, R varies with the coordinate y (interpolatation between $R = 7$ and $R = 3$, expressed in number of voxels).

- Distance d_{target} to the target point (i.e. the nerve-chiasm landmark). The penalty is proportional to this distance in order to force the centerline to immediately go towards the target point if other criteria do not give priority to some points. In particular when one part of the optic nerve has not been initially detected (negative voxels), the line should go in the direction of the target point. The associated cost is $c^{distance}_{(x,y,z)} = C^t d_{target}$ where C^t controls the importance of this cost. We fixed $C^l = 0.001$, to make it negligible compared to the previous criteria.

The cost of the node (x, y, z) is the sum of the three components: $c_{(x,y,z)} = c^{label}_{(x,y,z)} + c^{border}_{(x,y,z)} + c^{distance}_{(x,y,z)}$. The introduced cost determines the weights of edges in the graph. A directed edge between the point (x_1, y_1, z_1) and (x_2, y_2, z_2) has the weight of $c_{(x_2,y_2,z_2)}$. The shortest path between nodes corresponding to the two endpoints of the optic nerve is computed by Dijkstra's algorithm [14, 53]. The start point is the eye-nerve landmark as the optic nerve is generally well visible close to the eye. To the best of our knowledge, our approach is the first to combine deep learning with the search of the shortest path in a graph for segmentation of tubular anatomical structures. However, the idea of

computing optimal distances for segmentation of tubular structures appears in interactive level-set methods presented in [17, 11, 5]. The objective of these methods is to find a geodesic between two points in the image chosen by the user. The Eikonal equation is constructed based on voxel intensites and contrasts, and the problem is solved by Fast Marching [45], similar to Dijkstra's algorithm. Application of methods based only on image intensities may be difficult for segmentation of the optic nerves in MRI due, for instance, to the noise in images and local inhomogeneity of intensities within the optic nerve.

The final segmentation of the optic nerve is constructed from the centerline. As the optic nerve has a variable thickness, around each point (x, y, z) of the centerline we consider two spherical volumes $S^1_{(x,y,z)}$ and $S^2_{(x,y,z)}$ with associated radii $R_1 \leq R_2$. All voxels within $S^1_{(x,y,z)}$ are classified positive (optic nerve). Voxels of $S^2_{(x,y,z)}$ which are not within $S^1_{(x,y,z)}$ are classified positive only if they were positive in the original segmentation. We fixed $R_1 = 2.5$ and R_2 corresponds to the radius R defined previously (large close to the eye, smaller close to the optic chiasm).
Finally, we apply mathematical morphology [52] to reduce false positives corresponding to structures which are 'attached' to the optic nerve and have a similar appearance. As these false positives are often connected to the correct segmentation by thin segments (Fig. 6), we apply the morphological opening with three 1D structing elements of size 2 in the three directions and we take the largest connected component.

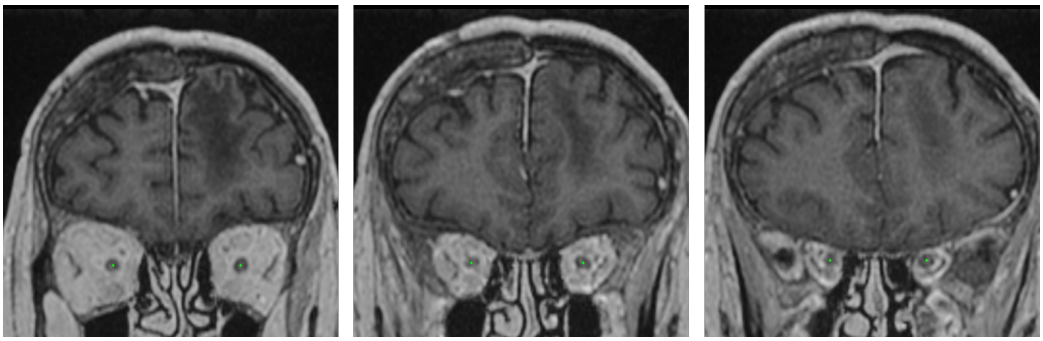

Figure 5: The centerlines of the optic nerves computed by our system on a test example (displayed on three different coronal slices). We assume one point of the centerline for each coronal slice between the two landmarks of the optic nerve.

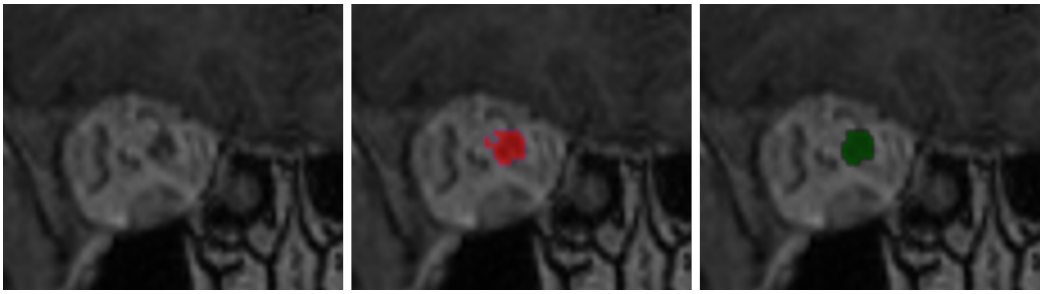

Figure 6: Use of mathematical morphology for reduction of false positives. Left: a coronal patch centered on an optic nerve. Middle: result obtained by the system on a test example without using mathematical morphology. Right: result obtained after application of morphological opening followed by taking the largest connected component.

3. Experiments

3.1. Data and preprocessing

We constructed a database of contrast-enhanced T1 MRIs acquired in the Centre Antoine Lacassagne (Nice, France), which is one of the three cancer centers in France equipped with proton therapy [30] systems.
The database contains 44 MRIs with provided segmentations of organs at risk and 50 non-annotated MRIs. The annotated images are used for training and cross-validated quantitative evaluation. For each scan, the ground truth segmentation was provided only for a subset of classes. The numbers of available segmentations for each class are reported in Table 1. The images without annotations are used for qualitative evaluation by a radiotherapist, as described in section 3.4.

The images were originally provided in Dicom format [33] and were heterogenous in terms of image intensities and geometrical properties such as size, spatial resolution and the visible part of the head. The ground truth segmentations were the ones used for routine radiotherapy planning and were provided in Dicom RT-Struct files [28], representing coordinates of polygons corresponding to contours of anatomical structures.

In order to use these images, we performed the following preprocessings. We used 3D Slicer [19] and its extension SlicerRT [40] to generate 3D volumes (in *nrrd* format) from Dicom slices and to generate binary label masks from RT-Struct files. All images were resampled to the same spatial reso-

lution, 0.7x0.7x0.9 (isotropic in axial slices, 0.9 spacing between slices) and then resized to dimensions 320x365x200. The 200 axial slices start from the top of the head, i.e. if an image originally has more than 200 axial slices, the bottom slices (close to the neck) are ignored. However, the input images had generally around 150 axial slices and the bottom slices were filled with zeros. To approximately normalize the image intensities, first we compute the maximum of an image, which is likely to be reached by a point on the fat or contrast-enhanced blood vessels. Then, all voxel values are divided by the value of the maximum and multiplied by a fixed constant.

3.2. Metrics for quantitative evaluation

To quantitatively evaluate our system, we perform 5-fold cross-validation on the set of 44 annotated MRIs. In each fold, 80% of the database is used for training and 20 % is used for test. For each class of interest, two results are reported. First, we report results obtained with our model trained on axial slices (denoted 'U-Net multiclass, axial' in the following), i.e. the raw output of the neural network, without postprocessing. Then we report results obtained after majority voting and postprocessing (denoted 'Final result' in the following).

The first metric we use is the Dice score, which measures the voxelwise overlap between the output and the ground truth segmentation. An important limitation of this metric is that it gives the same importance to very close and very distant mismatches. As the ground truth is often uncertain and

Table 1: Numbers of provided ground truth segmentations for different classes (in the database of 44 MRIs).

	Number of segmentations
Hippocampus	39
Brainstem	39
Eye	41
Lens	34
Optic nerves	40
Optic chiasm	41
Pituitary gland	29
Brain	37

noisy close to the boundaries of structures, the Dice scores are generally considerably lower for small structures. This is why, in addition to raw Dice scores, we also report results (Dice, sensitivity, specificity) obtained when a margin of one voxel is allowed, i.e. ignoring mismatches on the borders of the ground truth. This assumption means that a false positive on a voxel (x, y, z) which is directly neighboring with the ground truth segmentation is ignored, i.e. it is neither counted as false positive nor true positive. Similarly, a false negative (non-detection) on the border of the ground truth is ignored. The second used metric is the undirected Hausdorff distance expressed in millimeters (the coordinates of points are expressed in real values). The Hausdorff distance measures the length of the farthest mismatch between the output and and ground truth (false positive or false negative). It is therefore useful to assess the consistency of the result, i.e. presence of very distant mistmatches. However, its limitation is that it only measures the value of the maximal distance and therefore one misclassified voxel is sufficient to considerably increase the Hausdorff distance.

Therefore, we also measure the mean distance between the output segmentation A and the ground truth B, defined as follows:

$$M(A, B) = \frac{1}{|A| + |B|} \left(\sum_{a \in A} \inf_{b \in B} d(a, b) + \sum_{b \in B} \inf_{a \in A} d(b, a) \right) \tag{1}$$

where d is the Euclidean distance.

3.3. Quantitative results

The mean distances between produced segmentations and the ground truth segmentation ranged from 0.08 mm (for the brain) to 0.69 mm (for the pituitary gland), as reported in Table 5. The results are variable across the different organs, according to their size, the number of ground truth segmentations available for training and the overall complexity of the segmentation task.

The Dice scores are usually higher for large anatomical structures such as the brain and the brainstem. In particular, the borders of the ground truth are usually very uncertain, which represents a problem for quantitative evaluation for smaller classes. In large classes, the border region is small compared

to the entire volume of the class and therefore the mismatches on borders do not cause large drops of the metric. The highest Dice score was obtained for the brain (Dice score of 96.8). The lowest performances were obtained for the pituitary gland (mean Dice of 58, mean distance of 0.69 mm between the output and the ground truth). Segmentation of the pituitary gland is particularly challenging as it is small and difficult to be differentiated from surrounding structures. Moreover, the pituitary gland was the class with the lowest number of training examples (29 annotated cases, i.e. around 23 training cases in each of the 5 folds).
To take into account the uncertain borders of the ground truth, we also reported Dice scores, sensitivity and specificity ignoring mismatches on the border of the ground truth, as described previously. As most mismatches between the outputs and the ground truth are on noisy borders of organs, there is a considerable difference between the raw Dice score (Table 2) and the Dice score with tolerance to one voxel (Table 3).
However, the measured Hausdorff distances (Table 4) are higher for large classes. The highest mean Hausdorff distance is observed for the brain, for which it is of almost equal to 1 cm.

The combination of neural networks (trained respectively on axial, coronal, sagittal slices) by majority voting improved almost all metrics. The improvements were particularly large for the Hausdorff distance (Table 4) and the mean distance (Table 5). We observe that the majority voting removes almost all distant false positives and yields more robust results than a raw output of one neural network. The results were subsequently improved by additional postprocessings.

Table 2: Mean Dice scores (5-fold cross-validation) obtained on a set of 44 MRIs. 'Final result' denotes the result obtained after majority voting and postprocessing.

	U-Net multiclass, axial	Final result
Hippocampus	69.2	71.4
Brainstem	88.1	88.6
Eye	88.3	89.6
Lens	55.8	58.8
Optic nerves and chiasm	63.9	67.4
Pituitary gland	53.6	58.0
Brain	96.5	96.8

The postprocessing of the eyes consisted in setting a lower bound on the physical volume of the output segmentation. This simple procedure allowed to remove false positives and decreased the mean Hausdorff distance from

Table 3: Mean Dice score (5-fold cross-validation), sensitivity and specificity with tolerance to one voxel (ignoring mismatches on the borders due to the uncertainty of the ground truth).

	Dice score	Sensitivity	Specificity
Hippocampus	88.2	92.7	85.0
Brainstem	95.1	95.5	95.6
Eye	97.5	98.3	96.8
Lens	82.1	88.2	78.4
Optic nerves and chiasm	91.1	96.2	87.1
Pituitary gland	79.7	83.3	77.5
Brain	98.6	98.0	99.4

Table 4: Hausdorff distances in millimeters (5-fold cross-validation). 'Final result' denotes the result obtained after majority voting and postprocessing.

	U-Net multiclass, axial	Final result
Hippocampus	42.1	6.9
Brainstem	45.5	7.8
Eye	75.9	3.0
Lens	31.0	3.7
Optic nerves and chiasm	76.7	6.3
Pituitary gland	52.5	4.6
Brain	30.4	9.8

Table 5: Mean distances in millimeters (5-fold cross-validation).

	U-Net multiclass, axial	Final result
Hippocampus	0.97	0.66
Brainstem	0.26	0.26
Eye	0.35	0.11
Lens	1.29	0.63
Optic nerves and chiasm	1.09	0.48
Pituitary gland	2.45	0.69
Brain	0.07	0.08

12.2 mm (result of the majority voting) to 3 mm.
The postprocessing of the optic nerve decreased the number of false positives and enforced connectivity between the eyes and the chiasm, as described in section 2.2.2. False positives are removed when they are either too far from the centerline, hyperintense in T1-weighted MRI (fat surrounding eyes) or are disconnected from the main connected component after application of morphological opening removing thin segments. The Dice score with one-voxel tolerance increased from 89.6 (result of the majority voting) to 91.1 (after postprocessing) for the optic nerves and chiasm. The raw Dice score increased from 66.3 to 67.4.
The postprocessing of the brain consisted in taking the largest connected component and filling the 'holes' of the segmentation in axial, coronal and sagittal planes. As these 'holes' are usually small compared to the whole volume of the class (occupying a large part of the image), the variation of the metrics is limited. The Dice score increased from 96.7 to 96.8 and the Hausdorff distance decreased from 10.2 to 9.8.

To the best of our knowledge, the only deep learning work for segmentation of organs at risk in MRI is the one proposed in [38] which reported cross-validated results (mean distances in mm) on a set of 16 MRIs . The authors used a model-based segmentation [18] combined with a neural network for detection of boundaries of anatomical structures. The results reported by the authors for the anatomical structures we also segment are: 0.608 mm for the brainstem, 0.563 mm for the eyes, 0.268 mm for the lenses and 0.41 mm for the optic nerves and chiasm. Overall, the ranges of mean distances are therefore comparable to the ours.

Examples of the output segmentations (comparison to the ground truth) for the hippocampus, the brainstem and the optic nerve are displayed on Fig. 7, 8 and 9.

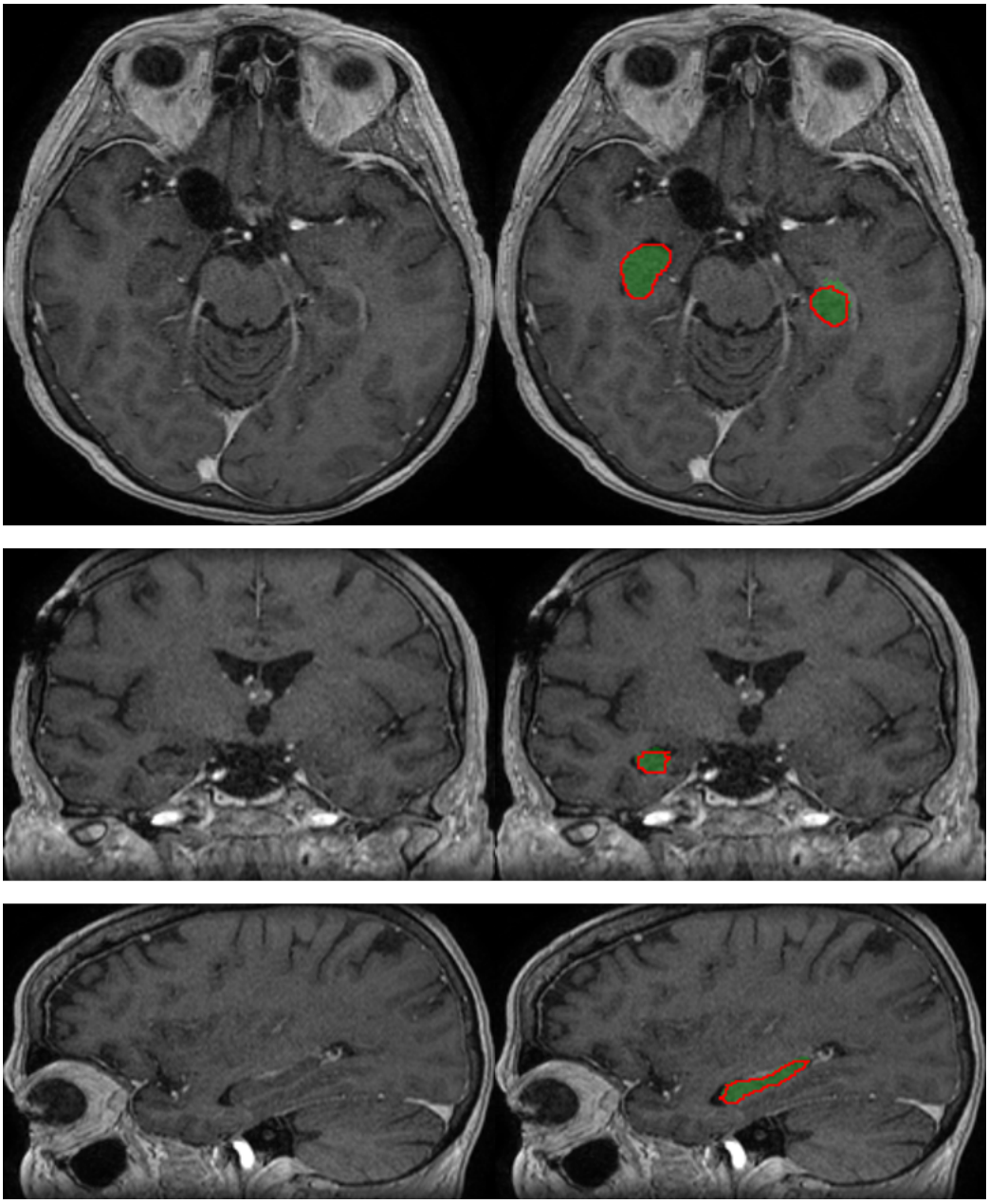

Figure 7: Segmentation of the hippocampus produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

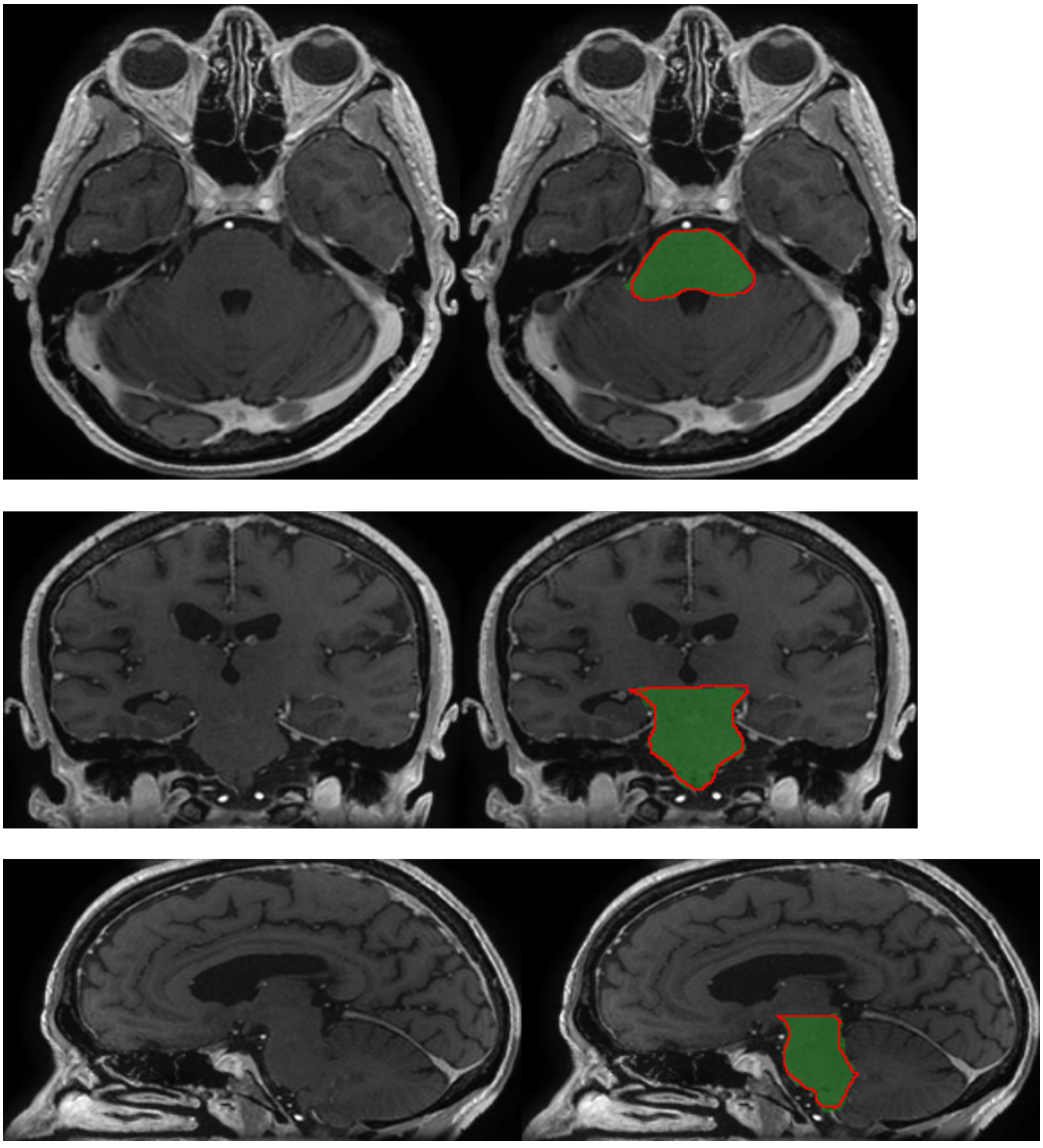

Figure 8: Segmentation of the brainstem produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

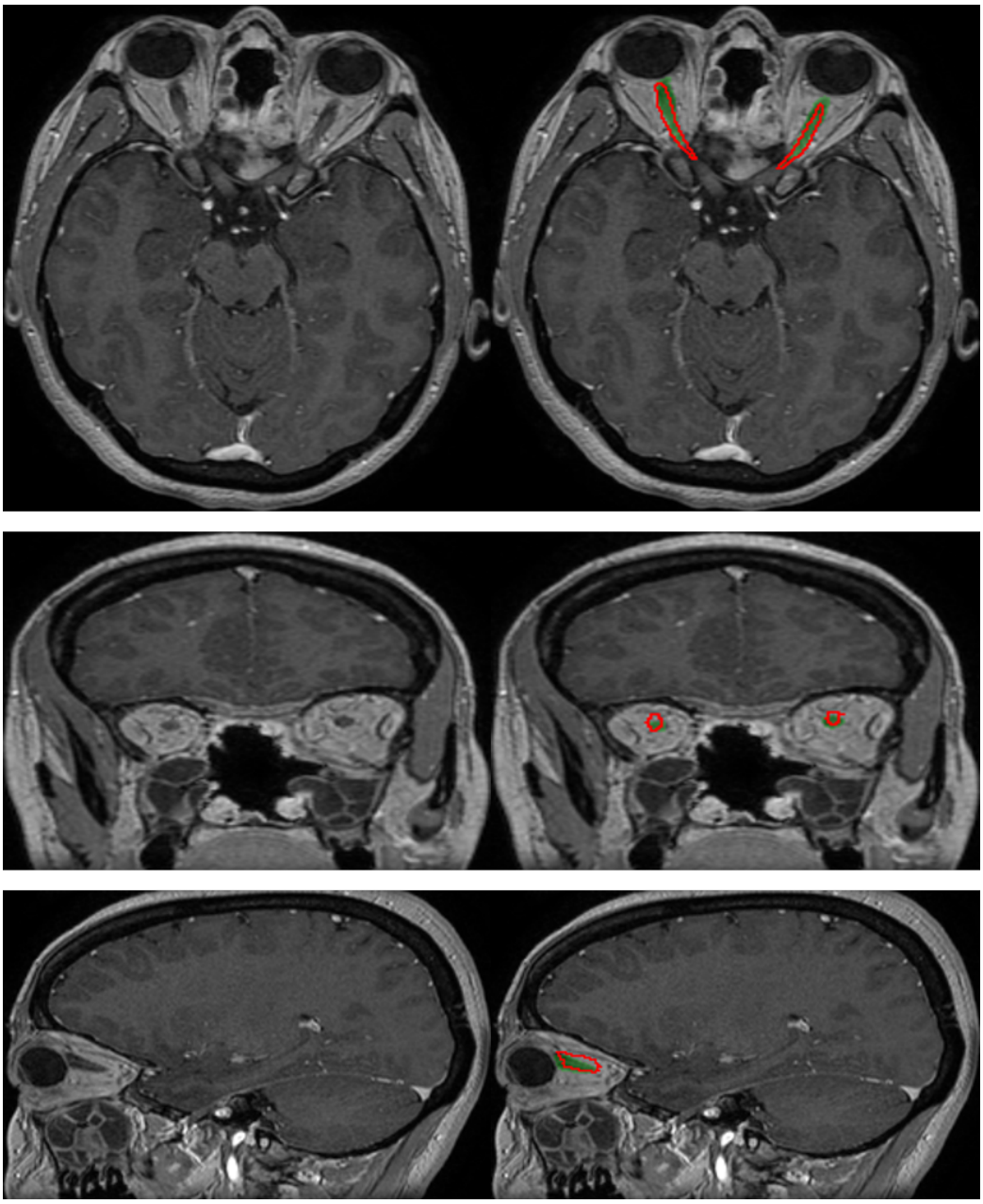

Figure 9: Segmentation of the optic nerves produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

3.4. Qualitative evaluation by a radiotherapist

The segmentations produced by our system on a set of 50 non-annotated MRIs are qualitatively evaluated by an experienced radiotherapist in order to assess their accuracy and utility for radiotherapy planning. For each of the 50 patients, the radiotherapist qualitatively evaluates the segmentations produced by our system for 12 anatomical structures (counting separately left and right components of bilateral classes), i.e. 600 segmentations are evaluated in total. The results are evaluated visually: no comparative dosimetric planning has been performed for corrected and uncorrected segmentations.. The segmentations are displayed with 3D Slicer [19]. Each of the 600 segmentations is assigned to one of the following categories:

- Accept: the radiotherapist would keep the segmentation for radiotherapy planning without any changes
- Accept, minor modifications: the segmentation is still acceptable for radiotherapy planning, i.e. some minor errors are observed but leaving out the modifications should not affect the dose distribution after the addition of safety margins of target volumes and organs at risk: planning target volumes (PTV) and planning organs at risk volumes (PRV)
- Accept, major modifications: the segmentation has necessarily to be corrected i.e. the resulting PTV and PRV differences before and after correction would have an important impact on the dose distribution (even if only few voxels are misclassified). The segmentation is however still good enough to be kept, i.e. it is less time-consuming to perform the necessary modifications than segmenting the structure from the beginning
- Reject: the segmentation has failed and keeping it would not save time compared to manually segmenting the structure from the beginning
- Not assigned: the structure is absent (e.g. organ removed by surgery) or invisible in the image because of a tumor

The results are summarized in Table 6. 73 % of the segmentations were assigned to the category *accept*, i.e. would be kept for radiotherapy planning without any modifications. Approximately 23 % of the segmentations were

Table 6: Clinical evaluation by a radiotherapist on 50 test cases.

	Accept	Accept, minor corrections	Accept, major corrections	Reject	N/A
Hippocampus left	39/50	8	1	2	0
Hippocampus right	45/50	5	0	0	0
Brainstem	22/50	26	1	1	0
Eye left	48/50	2	0	0	0
Eye right	45/50	4	1	0	0
Lens left	39/50	7	4	0	0
Lens right	42/50	6	2	0	0
Optic nerve left	44/50	6	0	0	0
Optic nerve right	40/50	10	0	0	0
Optic chiasm	19/50	26	4	1	0
Pituitary gland	19/50	25	3	0	3
Brain	36/50	14	0	0	0
Total	438/600	139	16	4	3

assigned to the second category, i.e. acceptable for radiotherapy planning but with recommendation to perform some minor corrections, usually on extremities of organs. The system produced therefore satisfactory segmentations in a large majority of cases. It was able to correctly delineate organs despite the important difficulties such as presence of tumors and the resulting mass effects, motion artifacts in MRI, different orientations of heads of patients and anatomical modifications resulting from previous surgeries undergone by the patient (removed tissues).

Segmentations of the eyes had the highest rate of immediate acceptation: 93 out of 100 segmentations were assigned to the *accept* category. The only segmentation which required a major modification was a case with a lesion inside the eye, possibly the polypoidal choroidal vasculopathy. The lesion was not classified by the system as part of the eye and therefore one part of the eye was not detected. The minor modifications recommended for other cases were generally to correct few non-detected voxels on the border of the eye (top or bottom axial slices) or few false positives on the anterior part of the orbit.
All segmentations of the optic nerves were found acceptable for radiotherapy

planning: 84 out of 100 segmentations were assigned to the *accept* category and the remaining 16 cases required only minor corrections. Most of the minor errors were non-detections for few voxels on the extremity of the optic nerve close to the eye (e.g. on the top axial slice). There was also at least one case of false positives on the neighboring arteries, close to the optic chiasm. Even if in the previous, quantitative evaluation, the metrics for the lenses were significantly lower than for other structures, most of their segmentations on the set of 50 MRIs were found satisfactory by the radiotherapist. Minor corrections were required in 13 out of 100 cases and major corrections were required in 6 cases. Most of the problems were non-detections, for instance observed in cases where the patient looks to the side and the system does not detect one side of the lens. The lenses are very small structures and their visibility is highly impacted by motion artifacts in MRI.

For the optic chiasm, the corrections were more frequently required but were usually minor: 19 out of 50 cases were assigned to the *accept* category and 26 cases required minor corrections. The minor errors were often false positives on the hypothalamus (the same issue was observed in the ground truth used to train the model) and sometimes on arteries neighboring the chiasm. The major corrections (4 cases) were mainly non-detections of a small subpart of the beginning of an optic nerve. In fact, even if only a small number of voxels is not detected (false negatives), the corrections are necessary as an excessive irradiation of one part of the optic nerve could make the entire nerve dysfonctional [25]. One segmentation was rejected due to non-detection of one part of the chiasm. This error appeared in a challenging case where the anatomy of the patient was modified by an important mass effect caused by a tumor.

Similar performances were obtained for the pituitary gland, located below the optic chiasm. Most of the minor (26 cases) and major (3 cases) required corrections correspond to non detections, typically on the 1-2 lowermost slices. In at least 2 cases, few false negatives were observed on the pituitary stalk (also observed in some ground truth segmentations used for training), which is the connection between the pituitary gland and the hypothalamus.

Even if segmentation of the hippocampus is difficult (low contrast with neighboring structures), in our evaluation it had one of the highest acceptation rates, with 84 segmentations in the *accept* category. However, it is also the only structure for which more than one segmentation was rejected. The two rejected segmentations correspond to cases where a large tumoral mass has grown near to the hippocampus, causing an edema having a similar intensity

in T1-weighted MRI. Morever, the tumors had a large necrotic core which may be confused with a ventricle by the system. In other cases, the required corrections (mostly minor) correspond usually to false positives (in particular on the amygdales, neighboring hippocampi and having a similar intensity in MRI T1) or some non-detections on the extremities of the hippocampus.
For the brainstem, 48 out of 50 segmentations were found acceptable for radiotherapy planning but required minor modifications in approximately half cases. The required corrections (false positives or non-detections) were almost exclusively on the uppermost axial slices (typically on 2 slices) which correspond to the top extremity of the brainstem. The only rejected segmentation corresponds to a case with a tumor adjacent to the brainstem and which was mistakenly included in the segmentation (false positives).
Finally, all segmentations of the brain (occupying a large part of the head) were found acceptable for radiotherapy planning even if they required minor corrections in almost one third of cases. The recommended corrections include, for instance, non-detections close to the cribriform plate (between the eyes) and false positives on bones.

In particular, we observe that the only two structures for which all segmentations were found acceptable for radiotherapy planning (without any major correction) are the ones for which a specific postprocessing was performed, i.e. the optic nerves and the brain.

4. Conclusion and future work

In this work we proposed a CNN-based method for segmentation of organs at risk from MR images in the context of neuro-oncology. The method was evaluated on clinical data.
First, we proposed a deep learning model and a training algorithm for segmentation of multiple and non-exclusive anatomical structures. The proposed methodology addresses problems related to computational costs and the variable availability of ground truth segmentations of the different anatomical structures (unsegmented classes). The neural network used in our method is a modified version of U-Net. The network is trained separately for segmentation in axial, coronal and sagittal slices. The three versions of the network are combined by majority voting.
Second, we proposed procedures to enforce anatomical consistency of the result in a postprocessing stage. In particular, we proposed a graph-based

algorithm for segmentation of the optic nerves, which are among the most difficult anatomical structures for automatic segmentation. The proposed postprocessings have shown their efficiency particularly in the qualitative evaluation by a radiotherapist. In particular, all segmentations of the optic nerves were found acceptable for radiotherapy planning.
The method was evaluated quantitatively on a set of 44 annotated MRIs, with 5 fold cross-validation and using several metrics. The segmentations produced by our system on a set of 50 non-annotated MRIs were qualitatively evaluated by an experienced radiotherapist. Despite the limited size of the training database (44 annotated MRIs) and the different challenges of the segmentation tasks (in particular, presence of tumors), a large majority of the output segmentations were found sufficiently accurate to be used for computation of irradiation doses in radiotherapy.

An important step of the future work is to adapt the method to multimodal data. Often, several types of images are acquired during radiotherapy planning for one patient, including CT scans and different MR sequences (T1, T2, FLAIR). Inclusion of different imaging modalities could improve segmentation of several structures but it comes also with new challenges related, for instance, to inter-modality registration and training of models on cases with missing modalities.
In our work, we used a variant of 2D U-Net to limit the GPU memory load. 2D CNNs have, however, the drawback of ignoring one spatial dimension. An interesting alternative would be to use 3D CNNs with anisotropic receptive fields such as the model proposed in [49].
As discussed in section 3, using the voxelwise overlap as the metric to compare the output segmentation with the ground truth has its limits, in particular due to uncertain borders of the ground truth. An interesting direction for future work would be to use a loss function penalizing distances between the output and the ground truth. However, computation of mean distances may represent important computational costs as it implies computation of distances for a large number of pairs of coordinates.
Even if the proposed postprocessing modules (hole-filling, thresholding on the size of connected components, inclusion of one class in another, graph-based algorithm to find centerlines) have been applied to specific anatomical structures, they can be adapted for segmentation of other structures. In particular, our graph-based algorithm could be used to compute centerlines of different tubular structures, such as blood vessels. The proposed post-

processings are not specific to MRI, except the intensity-based refinement of the segmentation of the optic nerve (hyperintensity in MRI T1 of the fat surrounding the optic nerve).
As for other segmentation tasks in medical imaging, availability of annotated training data is an important problem. Methods able to exploit weaker forms of annotations (bounding boxes, slice-level labels) for training of segmentation models are therefore of interest. In particular, methods combining weakly-annotated and fully-annotated training images were recently proposed in [35, 46]. As our system was able to produce accurate segmentations in a large majority of cases and the rare observed errors were mainly on boundaries of organs, the system could be used for generation of bounding boxes (subsequently verified by a human) which could be used to train segmentation models which are able to exploit this type of annotations.
Another important direction of the future work is to combine segmentation of organs at risk and segmentation of radiotherapy target volumes. In particular, a large variability of methods for tumor segmentation [36, 26, 49, 34, 39] were proposed in recent years. Deep learning could also be used for computation of irradiation doses [2] in radiotherapy planning.

Disclosures

The authors have no conflicts of interest to disclose.

Acknowledgements

Pawel Mlynarski is funded by the Microsoft Research-INRIA Joint Centre, France. This work was supported by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

References

[1] Alchatzidis, S., Sotiras, A., and Paragios, N. (2015). Local atlas selection for discrete multi-atlas segmentation. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 363–367. IEEE.

[2] Andres, E. A., Fidon, L., Vakalopoulou, M., Noël, G., Niyoteka, S., Benzazon, N., Deutsch, E., Paragios, N., and Robert, C. (2019). PO-1002 Pseudo Computed Tomography generation using 3D deep learning–Application to brain radiotherapy. *Radiotherapy and Oncology*, 133:S553.

[3] Argiris, A., Karamouzis, M. V., Raben, D., and Ferris, R. L. (2008). Head and neck cancer. *The Lancet*, 371(9625):1695–1709.

[4] Bauer, S., Wiest, R., Nolte, L.-P., and Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Physics in medicine and biology*, 58(13):R97.

[5] Benmansour, F. and Cohen, L. D. (2011). Tubular structure segmentation based on minimal path method and anisotropic enhancement. *International Journal of Computer Vision*, 92(2):192–210.

[6] Bloch, I., Colliot, O., Camara, O., and Géraud, T. (2005). Fusion of spatial relationships for guiding recognition, example of brain structure recognition in 3d mri. *Pattern Recognition Letters*, 26(4):449–457.

[7] Bondiau, P.-Y., Malandain, G., Chanalet, S., Marcy, P.-Y., Habrand, J.-L., Fauchon, F., Paquis, P., Courdi, A., Commowick, O., Rutten, I., et al. (2005). Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context. *International Journal of Radiation Oncology* Biology* Physics*, 61(1):289–298.

[8] Brosch, T., Peters, J., Groth, A., Stehle, T., and Weese, J. (2018). Deep learning-based boundary detection for model-based segmentation with application to MR prostate segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 515–522. Springer.

[9] Brouwer, C. L., Steenbakkers, R. J., van den Heuvel, E., Duppen, J. C., Navran, A., Bijl, H. P., Chouvalova, O., Burlage, F. R., Meertens, H., Langendijk, J. A., et al. (2012). 3D variation in delineation of head and neck organs at risk. *Radiation Oncology*, 7(1):32.

[10] Ciardo, D., Gerardi, M. A., Vigorito, S., Morra, A., Dell'Acqua, V., Diaz, F. J., Cattani, F., Zaffino, P., Ricotti, R., Spadea, M. F., et al. (2017). Atlas-based segmentation in breast cancer radiotherapy: evaluation of specific and generic-purpose atlases. *The Breast*, 32:44–52.

[11] Cohen, L. D. and Kimmel, R. (1997). Global minimum for active contour models: A minimal path approach. *International journal of computer vision*, 24(1):57–78.

[12] Commowick, O., Grégoire, V., and Malandain, G. (2008). Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiotherapy and Oncology*, 87(2):281–289.

[13] Commowick, O., Warfield, S. K., and Malandain, G. (2009). Using Frankensteins creature paradigm to build a patient specific atlas. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 993–1000. Springer.

[14] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms*. MIT press.

[15] Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., and Siddiqui, K. (2013). Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical image analysis*, 17(8):1293–1303.

[16] Criminisi, A., Shotton, J., Robertson, D., and Konukoglu, E. (2010). Regression forests for efficient anatomy detection and localization in CT studies. In *International MICCAI Workshop on Medical Computer Vision*, pages 106–117. Springer.

[17] Deschamps, T. and Cohen, L. D. (2001). Fast extraction of minimal paths in 3D images and applications to virtual endoscopy. *Medical image analysis*, 5(4):281–299.

[18] Ecabert, O., Peters, J., Schramm, H., Lorenz, C., von Berg, J., Walker, M. J., Vembar, M., Olszewski, M. E., Subramanyan, K., Lavi, G., et al. (2008). Automatic model-based segmentation of the heart in CT images. *IEEE transactions on medical imaging*, 27(9):1189–1201.

[19] Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al. (2012). 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic resonance imaging*, 30(9):1323–1341.

[20] Folk, M., Heber, G., Koziol, Q., Pourmal, E., and Robinson, D. (2011). An overview of the HDF5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*, pages 36–47. ACM.

[21] Fouquier, G., Atif, J., and Bloch, I. (2012). Sequential model-based segmentation and recognition of image structures driven by visual features and spatial relations. *Computer Vision and Image Understanding*, 116(1):146–165.

[22] Gauriau, R., Cuingnet, R., Lesage, D., and Bloch, I. (2015). Multi-organ localization with cascaded global-to-local regression and shape prior. *Medical image analysis*, 23(1):70–83.

[23] Ibragimov, B. and Xing, L. (2017). Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Medical physics*, 44(2):547–557.

[24] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[25] Källman, P., Ågren, A., and Brahme, A. (1992). Tumour and normal tissue responses to fractionated non-uniform dose delivery. *International journal of radiation biology*, 62(2):249–262.

[26] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78.

[27] Larsson, M., Zhang, Y., and Kahl, F. (2018). Robust abdominal organ segmentation using regional convolutional neural networks. *Applied Soft Computing*, 70:465–471.

[28] Law, M. Y. and Liu, B. (2009). DICOM-RT and its utilization in radiation therapy. *Radiographics*, 29(3):655–667.

[29] LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

[30] Levin, W., Kooy, H., Loeffler, J., and DeLaney, T. (2005). Proton beam therapy. *British journal of Cancer*, 93(8):849.

[31] Men, K., Geng, H., Cheng, C., Zhong, H., Huang, M., Fan, Y., Plastaras, J. P., Lin, A., and Xiao, Y. (2019). More accurate and efficient segmentation of organs-at-risk in radiotherapy with convolutional neural networks cascades. *Medical physics*, 46(1):286–292.

[32] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10):1993–2024.

[33] Mildenberger, P., Eichelberg, M., and Martin, E. (2002). Introduction to the DICOM standard. *European radiology*, 12(4):920–927.

[34] Mlynarski, P., Delingette, H., Criminisi, A., and Ayache, N. (2019a). 3D convolutional neural networks for tumor segmentation using long-range 2D context. *Computerized Medical Imaging and Graphics*, 73:60–72.

[35] Mlynarski, P., Delingette, H., Criminisi, A., and Ayache, N. (2019b). Deep learning with mixed supervision for brain tumor segmentation. *Journal of Medical Imaging*, 6(3):034002.

[36] Myronenko, A. (2018). 3D MRI brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer.

[37] Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., Askham, H., Romera-Paredes, B., Karthikesalingam, A., Chu, C., et al. (2018). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*.

[38] Orasanu, E., Brosch, T., Glide-Hurst, C., and Renisch, S. (2018). Organ-At-Risk Segmentation in Brain MRI Using Model-Based Segmentation: Benefits of Deep Learning-Based Boundary Detectors. In *International Workshop on Shape in Medical Imaging*, pages 291–299. Springer.

[39] Parisot, S., Wells III, W., Chemouny, S., Duffau, H., and Paragios, N. (2014). Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. *Medical image analysis*, 18(4):647–659.

[40] Pinter, C., Lasso, A., Wang, A., Jaffray, D., and Fichtinger, G. (2012). SlicerRT: radiation therapy research toolkit for 3D Slicer. *Medical physics*, 39(10):6332–6338.

[41] Ramus, L. and Malandain, G. (2010). Assessing selection methods in the context of multi-atlas based segmentation. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1321–1324. IEEE.

[42] Ramus, L., Malandain, G., et al. (2010). Multi-atlas based segmentation: Application to the head and neck region for radiotherapy planning. In *MICCAI Workshop Medical Image Analysis for the Clinic-A Grand Challenge*, pages 281–288.

[43] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

[44] Roth, H. R., Oda, H., Hayashi, Y., Oda, M., Shimizu, N., Fujiwara, M., Misawa, K., and Mori, K. (2017). Hierarchical 3D fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:1704.06382*.

[45] Sethian, J. A. (1999). Fast marching methods. *SIAM review*, 41(2):199–235.

[46] Shah, M. P., Merchant, S., and Awate, S. P. (2018). MS-Net: Mixed-supervision fully-convolutional networks for full-resolution segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 379–387. Springer.

[47] Tong, N., Gou, S., Yang, S., Ruan, D., and Sheng, K. (2018). Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical physics*, 45(10):4558–4567.

[48] Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., et al. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for

the Global Burden of Disease Study 2015. *The Lancet*, 388(10053):1545–1602.

[49] Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI Brainlesion Workshop*, pages 178–190. Springer.

[50] Wang, Y., Zhao, L., Song, Z., and Wang, M. (2018). Organ at Risk Segmentation in Head and Neck CT Images by Using a Two-Stage Segmentation Framework Based on 3D U-Net. *arXiv preprint arXiv:1809.00960*.

[51] Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903.

[52] Zana, F. and Klein, J.-C. (2001). Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation. *IEEE transactions on image processing*, 10(7):1010–1019.

[53] Zhan, F. B. and Noon, C. E. (1998). Shortest path algorithms: an evaluation using real road networks. *Transportation science*, 32(1):65–73.

[54] Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W., and Xie, X. (2019). AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics*, 46(2):576–589.